

# *Schnëssen*. Surveying language dynamics in Luxembourgish with a mobile research app

## 1 A mobile research app for Luxembourgish

Luxembourgish is a language with much variation, including regional varieties, socio-stylistic resources originating from French and German and ‘free’ variation that seems to be unconstrained in terms of social or linguistic categories. Many nouns, for example, allow two or more ways to indicate the plural, which cannot be traced back to regional, stylistic, or foreign language influence. These sources of variation, however, have only been researched to a limited extent, e.g. in relation to regional variation (in a comprehensive dialect atlas, i.e. *Luxemburgischer Sprachatlas* LSA 1963, using data from before WWII) or stylistic preference for co-occurring Germanic or Romance variants (see Conrad 2017).

Luxembourg is a trilingual country with French, German, and Luxembourgish as official languages in a complex sociolinguistic setting. Maintained by the educational system, a large share of the population can be regarded as trilingual on a rather high competence level. While French and German serve mainly as written languages and as lingua francas in the public sphere and at the workplace – 48% of the population are non-nationals – Luxembourgish has traditionally been used as a predominantly spoken language for most purposes by 50 to 70% of the total population (see Fehlen/Heinz 2016). With its increasing use as a written language, Luxembourgish is subject to ongoing processes of *Ausbau* and standardization and can serve thus as a prime example of a young, evolving and highly ‘plastic’ language (see Gilles, forthcoming).

Against this backdrop, the project *Schnëssen*<sup>1</sup> – *Är Sprooch fir d’Fuerschung* ‘Your language for research’ focuses on the surveying, analysing, and communication of variation in present-day spoken Luxembourgish. This includes traditional regional variants as well as contact-induced variation and other linguistically highly relevant phenomena. Methodologically, the project makes use of a dedicated mobile application developed in cooperation with the Swiss software studio ibros.ch. The technological backbone of the *Schnëssen* app is similar to prior applications used for other languages, focusing on recordings of audio data instead of written text. The *Schnëssen* app also offers sociolinguistic questionnaires in-app via an embedded mobile website. Projects such as this which aim to collect and analyse large datasets from app surveys are a relatively young pursuit within sociolinguistics (e.g. Leemann/Kolly/Britain 2018 and Britain/Leemann/Kolly 2018 for English dialects, Leemann/Kolly 2016 and Glaser/Bachmann/Hasse/Wanitsch 2018 for Swiss German and Hilton/Loerts/Visser/Jensma/Gooskens/Wanitsch/Leemann 2017 for Frisian). The *Schnëssen* app is similar in its focus on mass audio recordings, but is additionally capable of changing these recording items easily without requiring an update to the app. This feature allows for the creation of successive elicitation rounds. Furthermore, continuously publishing articles with

---

<sup>1</sup> *Schnëssen* is a Luxembourgish verb meaning ‘to chat, to gossip’.

results directly in the app helps to build a community of researchers and interested lay persons.

Given the availability of excellent network coverage, the ubiquity of smartphones in the Luxembourgish population<sup>2</sup> and the advanced technical specifications of contemporary smartphones (e.g. microphone quality), using a mobile app enables the collection of large amounts of data with comparatively little effort. In return, a lot of additional effort has to be expended in recruiting participants and disseminating results, e.g. via social media (see section 3). In conjunction with the linguistic purpose of this study, this project also seeks to establish long-term cooperation with a community of interested citizens which will make it possible to launch follow-up surveys or use the Schnëssen platform as a proxy for student and PhD projects. In doing so, this project seeks to combine different strands of sociolinguistic work, i.e. the analysis of variation and change from a variationist linguistics point of view, quantitative sociolinguistics (e.g. via questionnaires about language attitudes) and citizen science, i.e. the active participation of citizens in all aspects of project work. This article focuses on the design of the smartphone application and the methods of data collection (sections 2 and 3). Because of the recency of data collection in the project, it is not yet possible to provide detailed overall results. Sections 4 and 5 nevertheless present the overall structure of the corpus and exemplary results. The section 6 and 7 contain a discussion of the potential pitfalls of crowdsourcing in this context as well as an outlook.

## 2 Methods and design

The Schnëssen app has a modular design, offering different tasks in separate tiles (see *Figure 1* and the screencast video ‘Schnëssen\_Demo.mp4’ in the supplementary material). Before participants can take part in the recording or questionnaire tasks, they are asked to enter basic demographic information, such as place of birth, age, gender, first language, education level and language competencies. Participation is anonymous and no user registration is required to fulfil the tasks. Additionally, the app can handle multiple user profiles per device, to account for cases where e.g. a participant assists an elderly person to take part in the survey.

---

<sup>2</sup> In 2017, the number of mobile cellular subscriptions in Luxembourg per 100 people amounted to 136, higher than in France (106), Belgium (105) or Germany (134). See: <https://data.worldbank.org/indicator/IT.CEL.SETS.P2> [last access 24.7.2019]

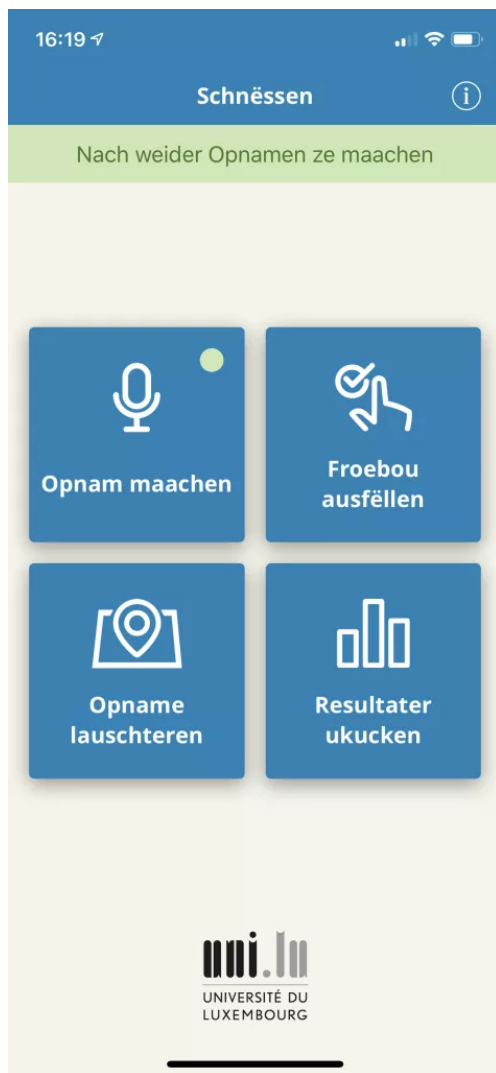


Figure 1: Screenshot of the main screen of the Schnëssen app

## 2.1 Collection of audio data

The main objective of the Schnëssen app is the documentation of a variety of linguistic phenomena contributing to the large and diverse variation of Luxembourgish. The linguistic phenomena focused on encompass all linguistic levels (phonetics/phonology, morphology, syntax, lexicon, pragmatics). Due to the large number of phenomena, and in order to avoid burdening participants with strenuous recording sessions, data collection is organised in ‘rounds’: each collection round is available for a couple of months and contains between 50 and 100 recording items. The fourth survey round was launched in April 2019. The first round (April to July 2018) focused on dialectological phenomena of Luxembourgish. Based on the dialect atlas from 1963 and further dialect surveys (e.g. Gilles 1999), a series of recording items were developed to trace regional dynamics in Luxembourgish over time and to help devise a new dialect atlas. Subsequent surveys focused on data collection for individual projects such as PhD theses, e.g. on issues of language standardisation or socio-pragmatic aspects of certain pronouns.

The main way in which participants engage with the app is through four different types of recording items: a translation task, a picture-naming task, a reading task and a question task. To offer the participants variety during recording sessions, these four types are mixed.

Some of the items regard ongoing public debates on language purism and decay (e.g. the words for ‘ant’, ‘hedgehog’ or ‘turtle’) in order to help participant motivation by making the material relevant and contemporary. The most important data to collect concerns the language used in the recording tasks. In order to access the entire range of linguistic variants, it is impossible to use written (Standard) Luxembourgish. Therefore, in the translation task, participants are asked to translate short sentences from German or French to Luxembourgish. Due to the high competence of most Luxembourgers in these two languages, these translations do not pose any difficulties. As an example for the translation task, consider the German sentence *Ihr fliegt vor Pfingsten nach Ägypten*. ‘You (pl.) are going to fly to Egypt before Pentecost.’ (see *Figure 2*). It contains (at least) five different phenomena of interest: phonological realisations for the pronoun *dir* ‘you’ ([di:ə / dɛ:ə]), the verb *fliegt* ‘fly’ ([flitt / flɪɪt / flɪkt]), the noun *Pfingsten* ‘Pentecost’ ([pæ:ɪftən / pæ:ɪstən / penʃtən / penʃstən] etc.), the preposition *nach* ‘to’ ([op / no]), and the country name *Ägypten* ‘Egypt’ ([e:ˈʒiptən / e:ˈʒyptən / ɛ:ˈgyptən] etc.). Using translations such as these, it is possible to survey a very large number of phenomena (more than 500 so far) within a very limited number of translation items.

Due to the typological closeness of Luxembourgish and German, we primarily use German as the default language for the translation items, except in cases where the priming effect might be too strong; where a German word is too close to its Luxembourgish counterpart, a French sentence is used instead. Nevertheless, especially in the case of lexical variation and due to language contact, priming effects for some items can only be avoided by using pictures. Hence, the picture-naming task is intended to collect Luxembourgish words for certain graphical representations of everyday items or concepts, e.g. the playing card shown in *Figure 2* (Luxembourgish *Joker* [ˈʒəʊkɐ / ˈdʒəʊkɐ] or *Stippi* [ˈftipi:]).

In the reading task, participants are asked to read a sentence containing linguistic phenomena that cannot be easily influenced by Luxembourgish orthography. In the example in *Figure 2*, this is the case for the ubiquitous word *Busarrêt* ‘bus stop’, which has been included in order to analyse word internal obstruent voicing, as in [ˈbusarɛ:] vs. [ˈbuzarɛ:].

Finally, in the question task, participants answer short questions in relation to specific lexical items or pragmatic aspects (e.g. ‘What is the word for the institution where convicted people are sent to?’ – *Gefängnis/Prison* ‘prison’; ‘How do you greet a foreign person?’ – *Moien/Salut/Zali/Bonjour*). This task is additionally designed in such a way to help detect a participant’s preference towards loan vocabulary from either French or German.

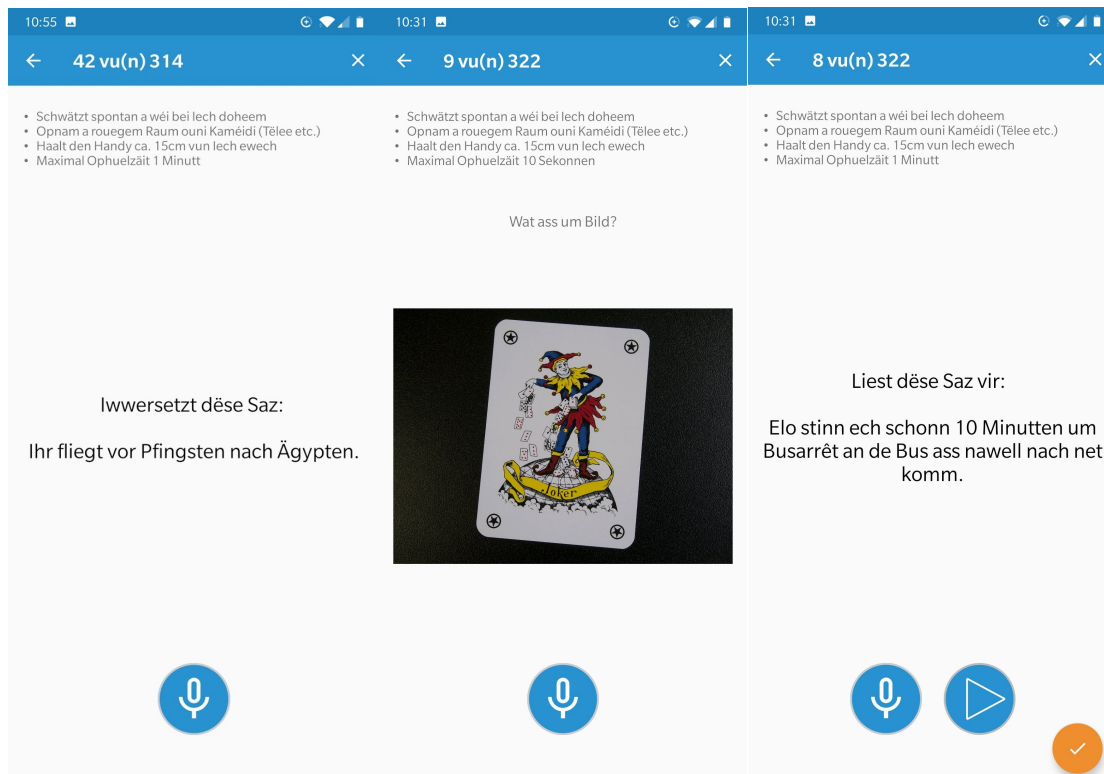


Figure 2: Screenshots of three recording tasks, from left to right: a translation task, a picture-naming task and a reading task

By clicking on the microphone button, participants can begin the audio recording for an item. The transfer of recorded audio data is done on a per-item basis, allowing the participant to interrupt and continue the recording session at any time. The sound files are recorded in uncompressed WAV-format, sampled at 44.1 kHz. For the analysis of lexical, morphological or syntactic aspects, virtually all recordings can be used. Concerning (articulatory or acoustic) phonetic investigations, some recordings have to be discarded due to bad sound quality, e.g. when recordings were made in a noisy environment such as in a car, on the bus or in a restaurant.

In general, this method of data elicitation has successfully produced results in line with the expected outcomes for the different phenomena. Given the nature of the unsupervised survey method, however, there is always the possibility that participants interpret the instructions differently than intended. This is predominantly the case when participants translate items using a divergent word order or alternative syntactic construction. Such instances aside, there has been very little misuse of the app so far (e.g. deliberately making wrong entries or unrelated recordings). The overwhelming majority of participants in fact report a rather serious disposition toward using the app. Many also evaluate the project very positively in a free feedback recording item at the end of the survey.

All content used for the different tasks is managed centrally via a server frontend. This makes it possible to add new items, reorder existing items, correct mistakes and to reactivate entire prior survey rounds if necessary or appropriate. Having this flexibility allows us to launch new survey rounds at any time, allowing us thus to be able to react to current events and trends in public discourse (see section 3). All sound files are stored on a dedicated server including a unique (anonymous) identifier to reference the files with the associated user profile. To facilitate (manual) analysis, recordings and user profiles are organised in *Google*

Spreadsheets, allowing collaborative work on the annotations. The screenshot in *Figure 3* illustrates one such annotation spreadsheet.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Eine Frau hat am Freitag die schwarzen Kleider genäht.						1: Freideg; Freides; 2: Freidj; 3: Freiden (Endung); 4: Freideg/Freides; 5: Freidj; 6: Freideg/Freides; 7: aner (a Remarken)													
2							1: gebitzt; 2: gebutt; 3: gebelzt; 4: gebout (mit Langvokal); 5: gebikst; 6: gebukt; 7: genäht; 8: genäht; 9: aner (a Remarken);													
3																				
4	pictu	rec	recordingUR	record	surveyID1	Ofspillen	Freitag	Remarken_F	genäht	Remarken_g	Variant_geni	Variant_Freit	RE_bilzen	Alter	Geschlecht	Höchsten Au	Kompetenz i	Kompetenz i	Aner Lokalle	Mutter
5	6	112	https://luxapssda	322	9	Ofspillen	7 Fraddig		4	gebout	Aner	RE	25 bis 34	Männlech	Fachhéischschou	7			5 KENG weider Af Jo	
6	6	112	https://luxapssda	424	17	Ofspillen	1		2	gebutt	Freideg/Freide RE	25 bis 34	Weiblech	Fachhéischschou	7			6 KENG weider Af Jo		
7	6	112	https://luxapssda	523	18	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Weiblech	Fachhéischschou	7			5 KENG weider Af Jo		
8	6	112	https://luxapssda	558	19	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Männlech	Fachhéischschou	7			5 Kanton Dikrech Jo		
9	6	112	https://luxapssda	610	20	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Weiblech	Fachhéischschou	7			5 Kanton Lëtzebu Jo		
10	6	112	https://luxapssda	712	23	Ofspillen	3		1	gebitzt	Freiden kee RE	25 bis 34	Weiblech	Fachhéischschou	6			4 Kanton Lëtzebu Jo		
11	6	112	https://luxapssda	814	24	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Weiblech	Fachhéischschou	7			6 KENG weider Af Jo		
12	6	112	https://luxapssda	957	25	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	35 bis 44	Weiblech	Fachhéischschou	7			7 Kanton Lëtzebu Jo		
13	6	112	https://luxapssda	1067	26	Ofspillen	3		2	gebutt	Freiden RE	25 bis 34	Weiblech	Fachhéischschou	7			5 Kanton Lëtzebu Jo		
14	6	112	https://luxapssda	1168	27	Ofspillen	7 Freddig		1	gebitzt	Aner kee RE	55 bis 64	Weiblech	1e (Lyée clai	7			7 Kanton Dikrech Jo		
15	6	112	https://luxapssda	1346	30	Ofspillen	1		2	gebutt	Freideg/Freide RE	25 bis 34	Weiblech	Fachhéischschou	7			7 An Däitschland q Jo		
16	6	112	https://luxapssda	1367	31	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	45 bis 54	Weiblech	COP/DAP (CATP	7			5 KENG weider Af Jo		
17	6	112	https://luxapssda	11223	36	Ofspillen	3		10	gebitzt	Aner	Freiden	35 bis 44	Weiblech	Fachhéischschou	7			4 An Däitschland q Neen	
18	6	112	https://luxapssda	1560	39	Ofspillen	3		1	gebitzt	Freiden kee RE	45 bis 54	Weiblech	COP/DAP (CATP	7			7 Kanton Esch-Uel Jo		
19	6	112	https://luxapssda	2735	116	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	35 bis 44	Weiblech	Fachhéischschou	7			7 Kanton Kappelen Jo		
20	6	112	https://luxapssda	2902	124	Ofspillen	3		1	gebitzt	Freiden kee RE	25 bis 34	Weiblech	Fachhéischschou	7			7 Kanton Kierf Jo		
21	6	112	https://luxapssda	2982	126	Ofspillen	3		2	gebutt	Freiden RE	25 bis 34	Weiblech	Fachhéischschou	7			6 KENG weider Af Jo		
22	6	112	https://luxapssda	2964	127	Ofspillen	3		1	gebitzt	Freiden kee RE	35 bis 44	Männlech	Fachhéischschou	6			5 KENG weider Af Jo		
23	6	112	https://luxapssda	3016	128	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Männlech	Fachhéischschou	7			6 KENG weider Af Jo		
24	6	112	https://luxapssda	3071	129	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	55 bis 64	Weiblech	Fachhéischschou	7			7 Kanton Esch-Uel Jo		
25	6	112	https://luxapssda	3185	130	Ofspillen	2		1	gebitzt	Freidig kee RE	35 bis 44	Weiblech	Fachhéischschou	7			6 Kanton Dikrech Jo		
26	6	112	https://luxapssda	3243	131	Ofspillen	3		2	gebutt	Freiden RE	25 bis 34	Weiblech	1e (Lyée clai	7			4 Kanton Esch-Uel Jo		
27	6	112	https://luxapssda	10363	132	Ofspillen	3		1	gebitzt	Freiden kee RE	35 bis 44	Weiblech	Fachhéischschou	6			3 An Däitschland q Jo		
28	6	112	https://luxapssda	3452	133	Ofspillen	3		1	gebitzt	Freiden kee RE	55 bis 64	Weiblech	13e/14e (Lyée	7			5 Kanton Esch-Uel Jo		
29	6	112	https://luxapssda	3460	135	Ofspillen	2		2	gebutt	Freidig RE	45 bis 54	Weiblech	COP/DAP (CATP	5			6 Kanton Réiden Jo		
30	6	112	https://luxapssda	11234	136	Ofspillen	3		1	gebitzt	Freiden kee RE	25 bis 34	Weiblech	Fachhéischschou	7			5 Kanton Esch-Uel Jo		
31	6	112	https://luxapssda	3603	138	Ofspillen	1 aner Reiefolleg		1	gebitzt	Freideg/Freide kee RE	25 bis 34	Männlech	13e/14e (Lyée	7			6 Kanton Esch-Uel Jo		
32	6	112	https://luxapssda	6321	139	Ofspillen	2		4	gebout	Freidig RE	25 bis 34	Männlech	Fachhéischschou	7			7 Kanton Iechtern Jo		
33	6	112	https://luxapssda	3668	141	Ofspillen	0 kee Saz		0 kee Saz	Näischt	Näischt	Näischt	65+	Männlech	Fachhéischschou	7			5 KENG weider Af Jo	
34	6	112	https://luxapssda	3794	144	Ofspillen	1		1	gebitzt	Freideg/Freide kee RE	35 bis 44	Weiblech	Fachhéischschou	6			6 Kanton Lëtzebu Jo		
35	6	112	https://luxapssda	3826	146	Ofspillen	3 aner Reiefolleg		1	gebitzt	Freiden kee RE	25 bis 34	Weiblech	13e/14e (Lyée	6			4 Kanton Lëtzebu Jo		
36	6	112	https://luxapssda	4188	147	Ofspillen	5		2	gebutt	Freidig RE	35 bis 44	Männlech	1e (Lyée clai	5			5 Kanton Dikrech Jo		

Figure 3: Screenshot of an annotation table in a Google Spreadsheet

## 2.2 Sociolinguistic questionnaire

Luxembourg's sociolinguistic make-up, with its complex relationships between Luxembourgish, French and German, is a prevailing topic in the public discourse, be it in the news, in private or on public transport. In order to contextualize the linguistic choices, participants make in the recording task and to gain a deeper understanding of language awareness, preferences, ideologies and attitudes, a sociolinguistic questionnaire is included in the Schnëssen app. The questionnaire is designed following a quantitative approach, asking participants to evaluate short statements on five-point Likert scales or in forced-choice questions (see *Figure 4*). In doing so, different aspects of the participants' perceptions and evaluations of everyday social practices can be addressed, such as language choices in typical everyday situations, assessments of linguistic and cultural diversity in Luxembourg, personal norm horizons in relation to correctness, and social positionings vis-à-vis language policy and change in Luxembourgish (for a general overview of attitudes towards multilingualism in Luxembourg, see Fehlen 2009; for the theoretical basis of this quantitative approach to attitudes, see Purschke 2015). In a subsequent analysis, the results of this study will be correlated with the participants' speech production in the recordings tasks. The questionnaire is hosted on a *LimeSurvey* server (LimeSurvey 2003ff.), which can be embedded directly into the app via a mobile website. This allows for flexible administration, updating and even replacement of the survey entirely without needing to update the app itself.

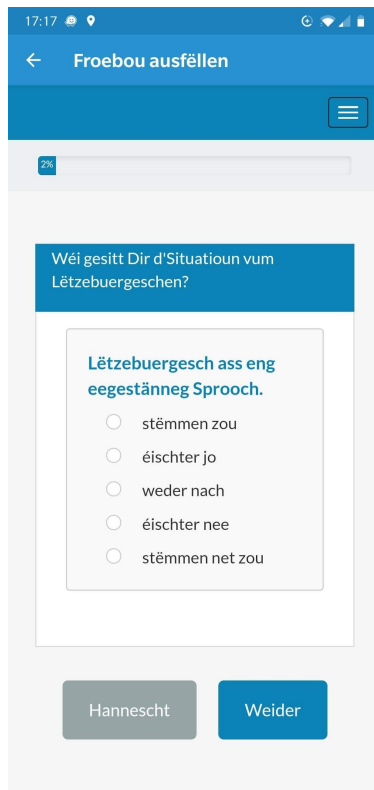


Figure 4: Screenshot of a forced-choice survey question in the sociolinguistic questionnaire

## 2.3 Presentation of data and results

All audio recordings are accessible to the public via an in-app map (see *Figure 5*). Zooming in and out reveals all the locations for which recordings have been submitted, with the number in the circles indicating the number of participants per location. Selecting a location opens a list of all items that have been recorded in this location. Selecting an item from this list then further displays the list of individual recordings together with the age groups of the respective participants and a time stamp of the recording (see *Figure 5*).



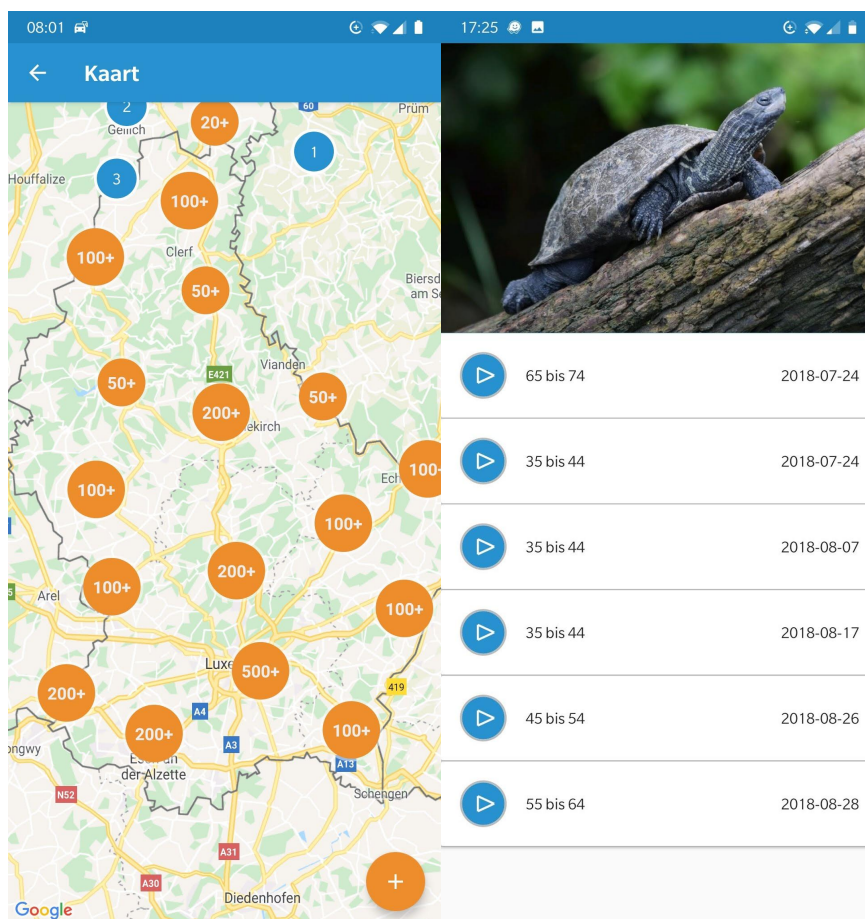


Figure 5: Screenshot of audio recordings on a map and as a list

Selected survey results are published in the results section of the app to inform participants about project progress and to demonstrate how the submitted data are being processed. This feature distinguishes Schnessen from similar apps, which do not tend to share their results so publicly. Shorter results articles analyse specific variation phenomena which could be of interest to the wider public and provide visualisations of quantitative aspects, e.g. frequencies of linguistic variants and distribution by age group, education level or gender. Variation phenomena with a regional distribution are presented in linguistic maps. The example article in *Figure 6* (left) shows the results of a survey question on the active language of the participants' mobile phones. *Figure 6* (right) illustrates the regional variants for the negation particle *net* 'not' ([nɛt], [nit], [nik], [nek]). To best cater to and attract the Luxembourgish public, all articles and in-app texts are written entirely in Luxembourgish. In addition, the results are published on the project's website <https://infolux.uni.lu/schnessen> as well as on our social media accounts.<sup>3</sup>

<sup>3</sup> Facebook: <https://www.facebook.com/Schnessen>, Twitter: <https://twitter.com/schnessen>.



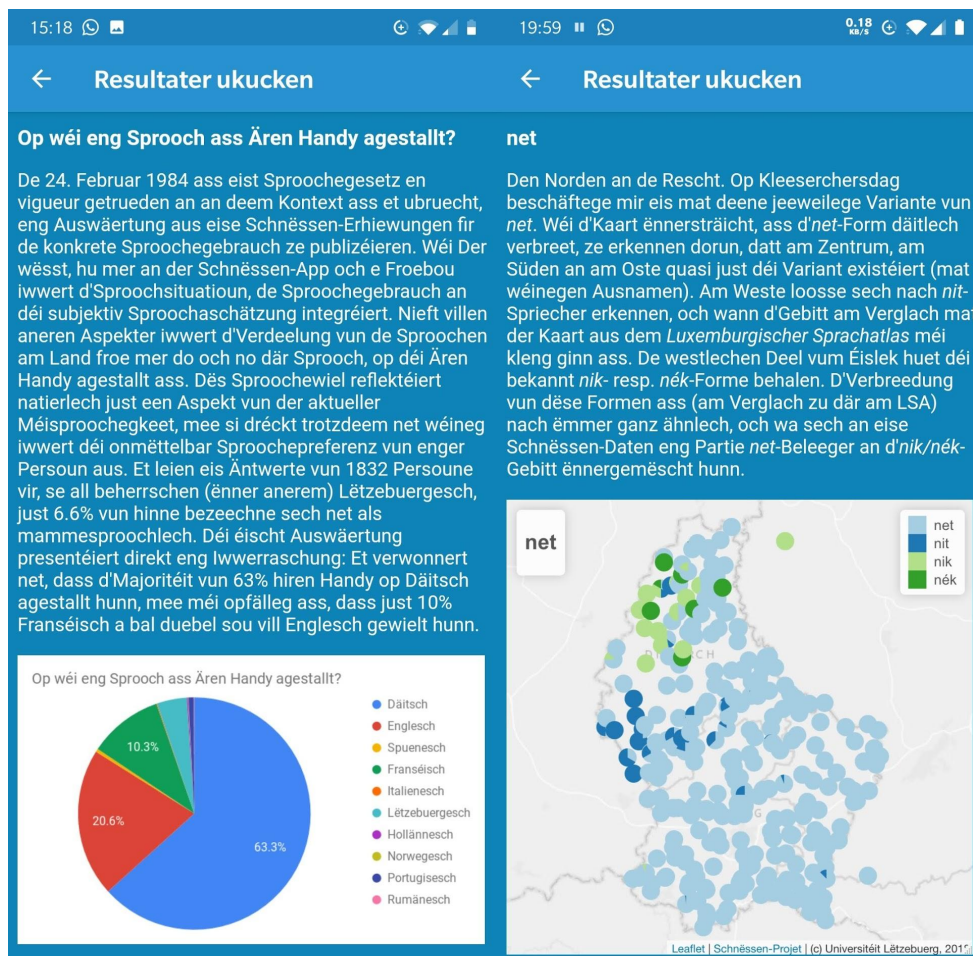


Figure 6: Screenshots of the result section

### 3 Participant recruitment and community building

In order to draw the public's attention to the app, recruit participants and build a community of linguistically interested citizens, we employ a variety of communication channels and strategies.

Firstly, we set up a Facebook app page (currently at 1200 likes), the most widely used social network in Luxembourg. Since the app release in April 2018, Facebook has been used to keep the participants informed about the app in general, to share news about recently added tasks and to share the processes and results of data collection and analysis. This helps to maintain public interest and motivation and functions as a direct method of interaction. This page also functions as a reward mechanism for the community, illustrating the direct outcome of their contributions.

To initiate interaction, those who have liked the Facebook page are asked to vote on which results analysis should be published next. Special news items are also used as a topical starting point for outreach activities whenever possible, e.g. results for the phonological variation of *Fussball* 'football' during the 2018 FIFA World Cup. This is intended to not only motivate potential participants to contribute to the survey but also to generate media attention (see section 3.2).



Figure 7: Screenshots of the Facebook app page

Two of the core aims of this Facebook page are to try to encourage the interested public to engage with our analysis of a specific linguistic phenomenon while at the same time trying to attract people to use the app. While the Facebook page has created a lot of interaction through comments, shares and likes, only a slight increase in the participation rate in the app itself can be observed as a direct consequence of this kind of social media campaigning. The same holds true for placing (paid) ads for the app on Facebook. Other media outlets have proven to be much more effective in this regard.

The 'traditional' media outlets – newspaper, radio and television – have become an important means for reaching the broader public. The app has been presented through interviews in print media and on a dedicated online science platform for the general public (science.lu). Most importantly, the app has been promoted through interviews and reports on RTL (radio, television, news portal, mobile application), the media group with the highest outreach in Luxembourg. Unsurprisingly, the collaboration with RTL had the greatest impact on app usage. RTL has published articles about the app in general, updates, new content and results. In a twelve-part 'summer series' during the summer holidays of 2018, app results were published online twice a week over a period of six weeks. These contributions were also briefly teased on RTL Radio. Taken together, the online articles have attracted a large number of participants to submit new recordings. The success of the cooperation with RTL also becomes apparent in *Figure 8*, depicting the evolution of the number of recordings per day: peaks in submission mostly correlate with instances of media coverage. Even if this connection is not always directly traceable, our survey proves that regular media campaigning can have a positive effect on the number of participants.

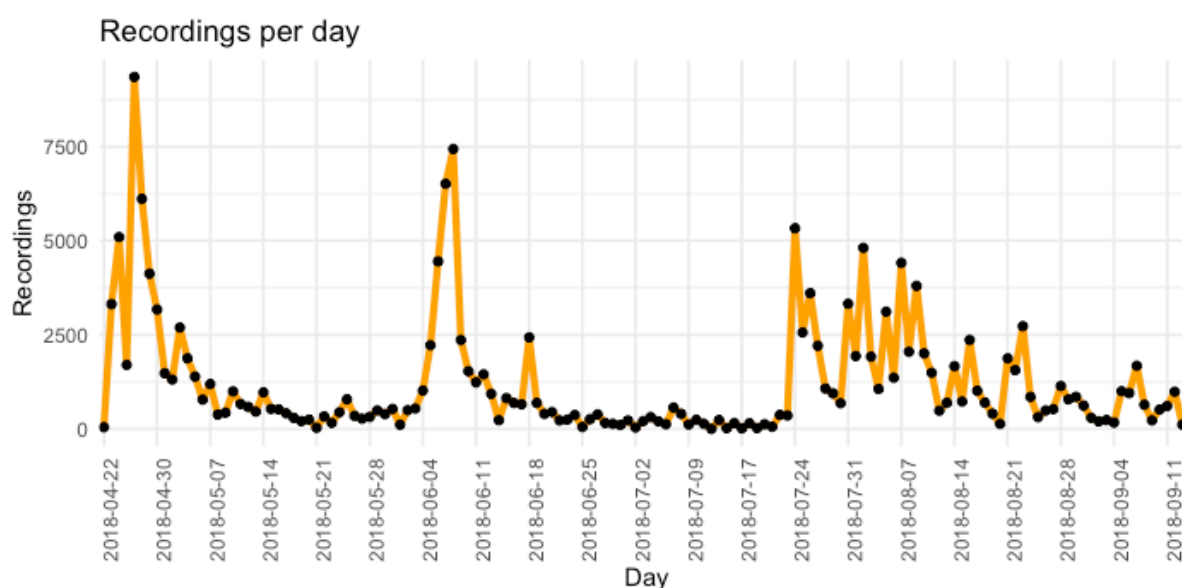


Figure 8 Received recordings per day in the Schnëssen app from 22.04.2018 to 11.09.2018

Another successful outreach activity was a citizen science workshop where the participants could research the collected data themselves together with the project team. 13 participants were introduced to the project and were trained to analyse and annotate recordings and to create maps and graphs. The focus of this workshop was not only the collaborative analysis of data but also the active involvement of participants in research activities. Giving the participants an insight into the daily work of a linguist and enabling them to pursue their own research interests allowed us to strengthen our ties with the participant community.

## 4 Corpus statistics

Due to the high awareness of language-related topics in the Luxembourgish speech community and the success of our media campaigning, participant recruitment and motivation has turned out to be a relatively easy task. Since its release in April 2018, the app has been downloaded 7900 times (Android: 2700, iOS: 5200) and is currently installed on 2500 devices (Android: 1200, iOS: 1300), which is a good user retention rate compared.

Across the four survey rounds of audio tasks, which took place between April 2018 and March 2019, we have received more than 210.000 individual recordings from nearly 3500 speakers in total (see *Table 1*). The number of items to record per round varied between 56 and 100, totalling 330 across all four rounds. A complete recording session for all items in a given round takes approximately 20 to 30 minutes to complete. The length of all audio recordings collected so far totals approximately 180 hours (assuming 3 seconds per recording), which constitutes the largest corpus of spoken Luxembourgish in existence.

Table 1: Overall results of the app-based recording tasks

	Recording items	Range of participants per item	Recordings
Round 1 (April '18)	100	820 - 2100	110.000

Round 2 (July '18)	100	180 - 1370	76.600
Round 3 (October '18)	74	190 - 430	20.400
Round 4 (March '19)	56	150 - 230	9.200
<i>Total</i>	330		216.200

As is typical for this kind of survey, many participants only open the app once to get an impression and record a few of the first items, never then returning to or continuing with their submissions. Across all four rounds, our participants nevertheless made 70 recordings on average. Only a few participants have contributed more than 100 recordings. Purschke (forthcoming) identifies similar participation patterns for the participatory linguistic landscape research app *Lingscape*, distinguishing between *casual users* with a low participation level and *regular* and *power users*.

Representativeness in scientific data is difficult to quantify, even more so for heterogeneous data originating from unsupervised crowdsourcing tasks. The following numbers may nevertheless give an indication of the relative size of the collected data. Assuming a total community of approximately 320.000 speakers of Luxembourgish (see Fehlen/Heinz 2016), for the first round of data collection, and depending on the recording item, between 0.3% (820 participants) and 0.7% (2100 participants) of this population are ostensibly present in the sample. To reach a comparable participation rate in e.g. Germany it would be necessary to recruit between 258.000 and 602.000 participants (assuming a total of 86 million speakers). Over 2800 participants answered the sociolinguistic questionnaire, many of which also participated in the audio recording tasks.

As for the demographic profile of the participant population, roughly two-thirds are female. About 90% of all participants state that Luxembourgish is a first language, while 10% indicate that they have learned Luxembourgish as a foreign language. With respect to regional distribution, participants originate from virtually every location in Luxembourg, as well as from approximately 100 villages outside Luxembourg. For the 400+ locations in Luxembourg, most participants come from the densely populated urbanised areas in the centre and in the south, but also the lesser populated rural areas show fairly high participation rates in relation to the total number of inhabitants. The geographical reach of this data is thus country-wide. Distribution across age groups (see *Figure 9*) is fairly even until the age group 45 to 54. For the two oldest age groups, the participation rate is lower, a common observation in app-driven projects and crowd-sourced data.

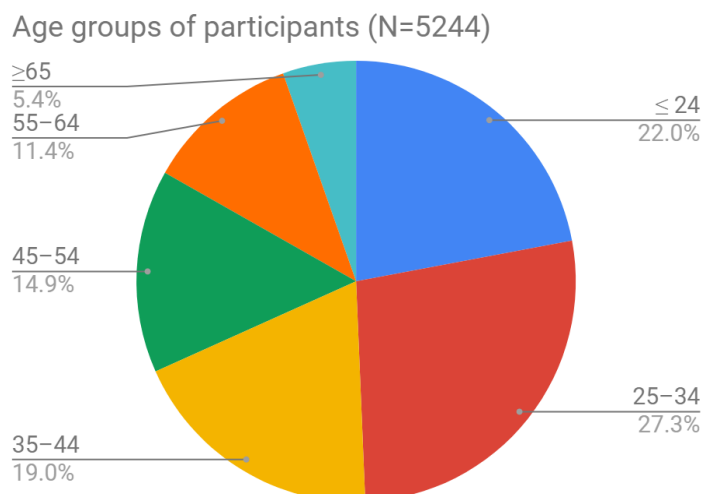


Figure 9: Distribution of participants per age group

## 5 First results

Due to the systematic composition of the recording tasks, the collected data covers the linguistic structure of Luxembourgish and its regional and sociolinguistic variation comprehensively, referring to the entire phonological system, large parts of morphology, selected aspects of syntax and pragmatics and numerous lexical items, among others. The dataset is sufficient to devise a new dialectological atlas, to study morphological and syntactic variation as well as (phonological, morphological and lexical) language contact with German and French. As data collection has only started recently, the following contains only initial findings.

The first results of this data have already been published in Gilles (2019), who investigates the ongoing merger of the fricatives [ç] and [j] for the minimal pair *frech* 'cheeky'/'Fräsch 'frog' using audio data from 1300 participants across all age groups. Due to this large database, it is possible to trace the spread of this merger through differing age groups. The high sound quality of the recordings also allowed for automatic phonetic segmentation with the MAUS system (see Winkelmann et al. 2017) and subsequent acoustic phonetic analysis. Martin (forthcoming) and Baumgartner/Busley/Fritzinger/Martin (forthcoming) study the socio-pragmatic constraints for the distribution of the two (feminine and neuter) personal pronouns used to refer to female persons, *si* 'she' and *hatt* 'it'. By simulating age differences and degrees of familiarity between female reference persons in various sentences in the recording tasks, the complex socio-pragmatic grounding and current change for these two pronouns can be clarified.

A further example of the potential of our data concerns phonological variation of the word *Freideg* 'Friday'. The regional distribution of variants for this word is well known from the dialectological atlas from 1963. Comparing the historical data to the variants we find in our dataset, it is possible to trace language change in Luxembourgish, e.g. the ongoing dialect levelling. The item has been embedded in a translation task and can be analysed based on approximately 1400 recordings (see Figure 10). To create the type of map pictured using pie charts containing the distribution of variants per location, the plotting functions of *R* (R Core Team 2019) were used. In our workflow for analysis, information is retrieved dynamically from data tables (see Figure 3), making the updating of maps and other visualisations easy



whenever new recordings arrive through the app. The HTML versions of these maps offer the possibility to zoom in on the map and to listen to the audio recordings of all locations by clicking on a given pie chart (see the HTML version ‘friday.html’ in the supplementary material). As can be seen from the juxtaposition of the two dialectological maps below, some striking changes have occurred: the northern area with the velarised forms *Fregdeg* [ˈfrægdə] (red), *Fregdig* [ˈfrægdɪ] (orange) seems to still be stable today, but has decreased in size when compared with the old atlas; the variant *Freddeg* [ˈfrædə] (light orange), which was present in a larger north-eastern area, can hardly be found anymore; and most importantly, the variant *Freiden* [ˈfrɑɪdən] (dark green), which was documented previously only for a small region in the south-west, can be found in the entire country according to our data, and is now challenging the standard variant *Freideg* [ˈfrɑɪdə] (dark blue). This example thus underlines the ongoing restructuring of the Luxembourgish dialect-geography.

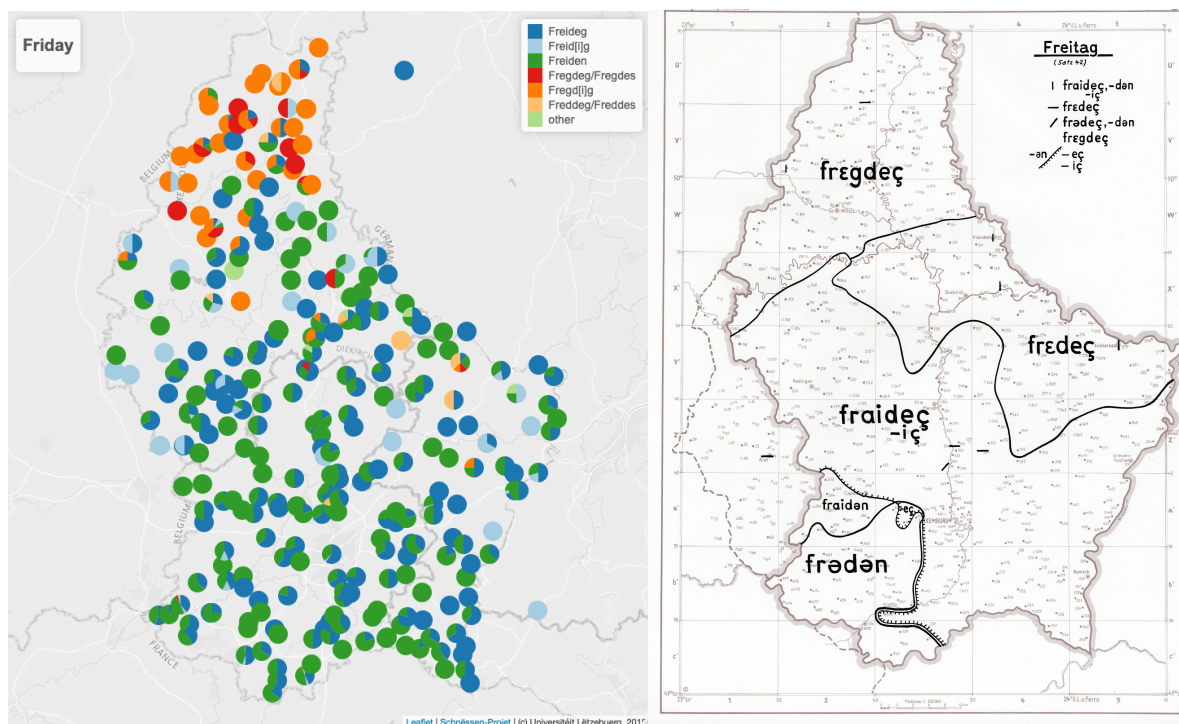


Figure 10: Regional variation for *Freideg* ‘Friday’: juxtaposition of the results for the present-day (left) and the historical situation (right; map 109 from LSA 1963)

The ongoing shift in the regional distribution of variants can be linked to changes across the age groups of the speakers. In brief, *Figure 11* depicts an apparent-time analysis, illustrating a correlation between the usage of the variant *Freideg* (dark blue) and older age, and an inverse correlation for usage of *Freiden* (dark green), i.e. over time *Freideg* is being replaced by *Freiden*. As expected, the change towards a new variant is mainly driven by younger people. The large dataset (here 1400 recordings) shows a rather homogenous distribution across the age groups and is therefore, in combination with the careful interpretation of further social parameters, suitable for apparent-time studies. This dataset has thus great potential in allowing for analyses of patterns of language variation and change in great quantitative detail.

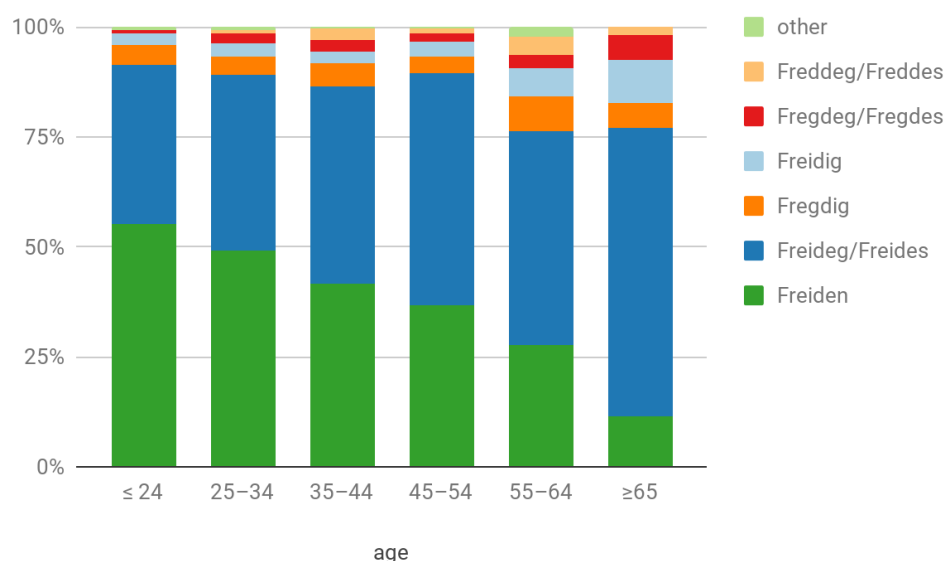


Figure 11: Apparent-time analysis for the variants of *Freideg* 'Friday' across six age groups

## 6 Potential pitfalls of crowdsourcing

Crowdsourcing, as has been shown in this study, is evidently an effective method for compiling a large audio corpus. For under-researched languages such as Luxembourgish, this becomes a crucial tool in achieving a comprehensive overview of present-day language variation. The results gathered through this process can also then facilitate in-depth studies on specific aspects of language variation and change.

There are however limitations to the collected data: the (rather classical) method of translating from one language to another is likely to generate more experimental than spontaneous language data and the structure of the source language (here German and French) may influence the resulting translation. This influence might be less pronounced for phonological or morphological phenomena, but can play a role for lexical or syntactical phenomena. Through careful design and testing of the translation tasks, these potential problems have been limited as far as possible.

A general drawback of nearly all crowdsourcing activities is that they rely on volunteers and might thus run into the risk of demographic bias: since there is hardly any reward mechanism for participation, it is likely to mainly reach participants who are intrinsically motivated to contribute data and to support research. Similarly, those who are typically less interested in aspects of language use or do not have easy access to today's digital infrastructures are more difficult to reach. In the case of the Schnëssen app, these factors led primarily to biases in gender (more women than men) and age (more younger speakers than elderly speakers), which must be taken into account in any research. While such a demographic bias is typical for technologically driven projects, there are ways to reduce the influence of said bias on corpus structure. For example, we are encouraging participants to (assist and) record their parents and grandparents. Apart from such limitations, we are able to collect data from a broad range of demographic groups including different education levels, dialect regions and language competencies.

Another demographic aspect to consider relates to the target audience of the app. While we are specifically targeting fluent speakers of Luxembourgish (by app design and



media campaigning), we would also like to collect data from speakers with Luxembourgish as a second language. Given the amount of detailed knowledge required, however, e.g. for the naming task and the high language requirements for the translation task, the app is likely to be prohibitive for many speakers and thus limits our potential audience to a subsample of the speech community that is likely to be Luxembourgish by nationality and trilingual. Thus, many aspects related to contact-induced variation (e.g. with Portuguese or Italian) or learner varieties of Luxembourgish are missing in the data set.

One of the main criticisms in relation to crowdsourced data relates to the quality and reliability of the collected data in general (see Lewandowski & Specht 2015). In our case, not only is the data collected with the Schnëssen app demographically biased, but we can also expect over-reporting of rare regional variants that are either not actively used by the speakers (“remembered forms”) or used specifically to demonstrate “Luxembourgishness” despite current developments in language change (“desirable forms”). Given the vast amount of phenomena and speakers from different demographic groups and locations, however, the dataset allows for a wide range of apparent-time analyses, helping to mitigate these problems. The results of the Schnëssen survey can thus provide invaluable insights into present-day variation and change in spoken Luxembourgish, especially in comparison with small-scale in-depth studies of specific locations or phenomena.

## 7 Outlook

The first year of the Schnëssen app has been very successful, with a high number of participants contributing recordings and completing questionnaires despite the time-consuming design of the different tasks. After the initial peak in participation at launch, there is now a degree of stagnation within our participant community. Although we are still receiving recordings and questionnaires daily, participation rate is far below the first few months after release. The sheer quantity and linguistic richness of the data collected so far is nevertheless (often more than) sufficient to study all originally compiled research questions. The next release of the app will contain a push notification function to directly inform participants about new recording items, questionnaires or results. By adding gamification elements (quizzes etc.), which are used for example in the *Gschmöis* app for Swiss German (Glaser/Bachmann/Hasse/Wanitsch 2018), participants might be better encouraged to take part.

Alongside this update, the app will be further developed into a research tool for PhD candidates and student projects, allowing them to access our pool of participants and to launch personal projects and experiments within the app. Other starting points for further analysis include (semi-)automatic segmentation and annotation of recordings using machine learning algorithms.

## Bibliography

Baumgartner, Gerda, Simone Busley, Julia Fritzingler & Sara Martin. forthcoming. *Dat Anna, et Charlotte und s Heidi*: Neutrale Genuszuweisung bei Referenz auf Frauen als überregionales Phänomen. Proceedings of the 6th Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen (IGDD) (ZDL Beihefte). 13.-15.09.2018, Marburg, Germany.

Conrad, François. 2017. *Variation durch Sprachkontakt. Lautliche Dubletten im Luxemburgischen* (Études Luxembourgeoises/Luxemburg-Studien 14). Frankfurt: Peter Lang.

Fehlen, Fernand. 2009. *BaleineBis: une enquête sur un marché linguistique multilingue en profonde mutation. Luxemburgs Sprachenmarkt im Wandel*. Luxembourg: SESOPI Centre intercommunautaire.

Fehlen, Fernand & Andreas Heinz. 2016. *Die Luxemburger Mehrsprachigkeit. Ergebnisse einer Volkszählung*. Bielefeld: Transcript.

Glaser, Elvira, Sandro Bachmann, Anja Hasse & Daniel Wanitsch. 2018. gschmöis - Die Smartphone-App zum Schweizerdeutschen. Mobile app for iOS and Android.

Gilles, Peter. 1999. *Dialektausgleich im Lëtzebuergeschen. Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Tübingen: Niemeyer.

Gilles, Peter. forthcoming. 40. Komplexe Überdachung II: Luxemburg. Die Genese einer neuen Nationalsprache. In Joachim Herrgen & Jürgen Erich Schmidt (eds.), *Language and Space - An International Handbook of Linguistic Variation. Vol. 4 Deutsch* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 30.4.). Berlin/New York: De Gruyter Mouton.

Gilles, Peter. 2019. Using crowd-sourced data to analyse the ongoing merger of [ɛ] and [ɪ] in Luxembourgish. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1590-1594). Canberra, Australia: Australasian Speech Science and Technology Association Inc. < [http://intro2psycholing.net/ICPhS/papers/ICPhS\\_1639.pdf](http://intro2psycholing.net/ICPhS/papers/ICPhS_1639.pdf)>

Hilton, Nanna, Hanneke Loerts, Willem Visser, Goffe Jensma, Charlotte Gooskens, Daniel Wanitsch & Adrian Leemann. 2017. *Stimmen: A Citizen Science App for Languages*. University of Groningen. Mobile app for iOS and Android.

Leemann, Adrian, Marie-José Kolly. 2016. Big Data for analyses of small-scale regional variation: A case study on sound change in Swiss German. In Christoph Draxler & Felicitas Kleber, *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum. P und P12*. München. <https://doi.org/10.5282/ubm/epub.29405>

Lewandowski, Eva, & Hannah Specht. 2015. Influence of volunteer and project characteristics on data quality of biological surveys. *Conservation Biology* 29/3. 713–723.

LimeSurvey GmbH. 2003ff. *LimeSurvey - An open source survey tool*. Hamburg: LimeSurvey Development Team.

Martin, Sara. forthcoming. *Hatt* or *si*? Neuter and feminine gender assignment in reference to female persons in Luxembourgish. In Antje Dammel & Corinna Handschuh (eds.), special issue of *Language Typology and Universals* (STUF).

Purschke, Christoph. 2015. REACT – A constructivist theoretic framework for attitudes. In Dennis Preston & Alexei Prihodkine (eds.), *Responses to Language Varieties: Variability, processes and outcomes*. Amsterdam/Philadelphia: John Benjamins, 37-54.

Purschke, Christoph. forthcoming. Using crowdsourced data to explore the linguistic landscape of cities. Results from the participatory research project *Lingscape*. In Stanley Brunn & Roland Kehrein (eds.), *Handbook of the Changing World Language Map*. Heidelberg: Springer.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Winkelmann, Raphael, Jonathan Harrington & Klaus Jänsch. 2017. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45. 392-410.